International Journal on Future Revolution in Computer Science & Communication Engineering
Volume: 4 Issue: 12
ISSN: 2454-4248
101 – 104

Taxonomy of Big Data Analytics: Methodology, Algorithms and Tools

Rakesh Kumar Singh
Scientist-E
G.B. Pant National Institute of Himalayan Environment and Sustainable Development,
Kosi-Katarmal, Almora-263643, Uttarakhand, India.
rksingh@gbpihed.nic.in

***Abstract*—** With the advent of revolution in the digital world, there is a rise in different type of data drastically. The data consists of structured and unstructured form of data, which is complex in nature and suppose to be stored, processed and finally analyzed to get some useful result for some organisation. As it is now not possible for traditional analytical methods to handle such a huge data, so a number of new and highly performance algorithms are developed for the efficient and useful analysis of big data. Analysis of these massive data requires a lot of efforts at multiple levels to extract knowledge for decision making. Therefore, big data analysis is a current area of research and development. This paper gives a brief introduction to big data and about the existing challenges, research, analytics tools, algorithms, open issues and future research direction for the study of big data analysts.

Keywords- *Data, Big Data, Database, Social Media, Analytics, Structured data, Unstructured Data.*

*****

I. INTRODUCTION

Data are generated from various sources and the fast transition from digital technologies has led to growth of big data. The term "Big Data" is not very new. Initially a terabyte was considered as Big Data. But as time passed, data kept on increasing. In today's time, many organizations are collecting, analyzing and storing a tremendous amount of data in order to complete a number of tasks. This large amount of data is commonly termed as "Big Data", as it is huge in amount. The term "Big Data" refers to the massive amounts of data that collect over time that are difficult to analyze and handle using common database management tools. Big Data includes e-mail messages, photos, business transactions, surveillance videos and activity logs, scientific data from sensors can reach mammoth proportions over time, and Big Data also includes unstructured text posted on the Web, such as blogs and social media etc. One perspective is that big data is different kinds of data in huge amount which cannot be easily handled by traditional relational database management systems (RDBMSs). Some people consider 10 terabytes to be big data, but any numerical definition is likely to change over time as organizations collect, store, and analyze more data. Another useful perspective is to characterize big data is having the concept of 3 V's: Volume, Velocity and Variety.

- ***Volume:*** Volume refers to the amount of data generated through websites, portals, web applications, online applications, etc. Volume encompasses the available data that are out and need to be accessed for decision making of some problem. Currently exponential growth can be seen in the data storage.
- ***Velocity:*** Velocity refers to the speed of data generation. The speed of generation and procession of data to meet the demands, determines the real potential in the data. The flow of data is massive and continuous. The Big Data velocity deals with the speed of data flow from sources like networks, social media sites, mobile sensors, mobile devices, GPS locations, application logs, etc.
- ***Variety:*** Data can be stored in multiple formats. For example excel, access, database, text files, etc. In Big Data, variety refers to all the structural and unstructured data that has the possibility of being generated or collected by users. Mostly structured data are in the form of texts, pictures, videos, tweets, etc. Emails, hand-written text, audio recordings fall in the category of unstructured data. It is all about the ability of classifying the data as incoming data into various categories.

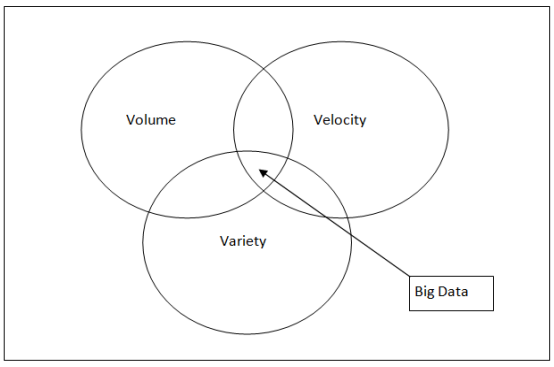

Fig.1. Taxonomy of Big Data.

II. SOURCE OF BIG DATA

Big Data has a number of sources. On every click of mouse on a website Web logs are captured and finally analyzed in order to create better understanding. Social media such as Instagram, Facebook, Twitter, etc. can generate tremendous amounts of posts. This all can be captures and analysed to create a new set of data for their audience. A tremendous

 ISSN: 2454-4248
101 – 104

amount of geospatial (e.g., GPS) data is gathered through different sources, e.g. mobile phones, which generate locations of each mobile user. Data related to Images, voice, and audio can be analyzed for applications in security systems.

III. BIG DATA ANALYTICS

It is the process of examining large and varied data sets, i.e., big data, to uncover hidden patterns, unknown correlations, market trends, customer preferences and other useful information that can help organizations make more-informed business decisions. Data analytics technologies and techniques provide a means of analyzing data sets and drawing conclusions about them to help organizations make informed business decisions. Big data is creating a decision support data management system in quite a new manner. The use of Analytics plays a key role in deriving values from big data. Collecting and storing big data creates little value; it is only data infrastructure at this point. It must be analyzed and the results used by decision makers and organizational processes in order to generate value.

IV. METHODOLOGY OF BIG DATA ANALYTICS

A research methodology can help big data managers collect better and more intelligent information. In terms of methodology, big data analytics differs significantly from the traditional statistical approach of experimental design. Data is the base of analytics. The objective behind this approach is to predict the response behaviour of the input data. Here various stages of lifecycle of Big Data Analytics have been discussed:

- ***Data identification and collection****:* This is the first phase of the methodology, identification of a large number of data sources have to be identified depending upon the severity of problem. Number of data resources gives the chances of finding core mean more chances of finding hidden correlations and patterns. More the data more chances of finding. Tools are required for capturing keywords, data and information from these heterogeneous data sources.
- ***Data storage****:* Every sort of data is to be stored in databases/ data warehouse. SQL databases are not needed to work in Big Data. Various frameworks and databases developed by organizations like Apache, Oracle etc. that allow analytics tools to fetch and process data from these repositories.
- ***Data filtering and noise elimination****:* In this phase, removal of replicates, corrupt and irrelevant data objects from gathered information takes place.
- ***Data classification and extraction****: Here,* extraction of incongruent data and conversion of it to a common data format takes place by using analytics tool. Extraction of relevant fields or texts also takes place here to reduce the volume of data.
- ***Data cleansing, validation and aggregation****:* In this phase, validation of rules based on the business case takes place to confirm the necessity and relevance of data extracted for analysis. Aggregation helps in combining multiple data sets to fewer based on common attributes.
- ***Data analysis and processing***: Data mining and analysis takes place in this phase to establish unique and hidden patterns for making business decisions.
- ***Data visualization****:* This phase involves visual or graphical representation of analysis results to make it easier to understand.

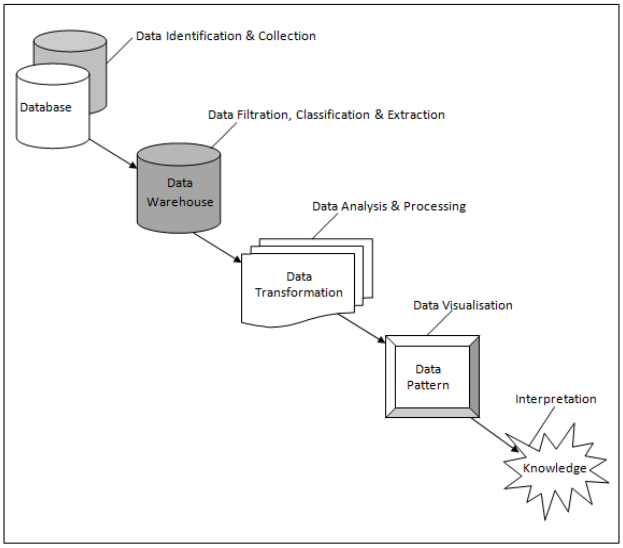

Fig.2. Various stages of Big Data Analytics.

V. BIG DATA ANALYSIS ALGORITHMS

Data mining algorithms for data analysis plays a very important role in the process of big data analysis in terms of memory requirement, computational cost, and accuracy of the results. Here a brief description of search and analysis algorithms is given, explaining the importance of big data analysis.

- ***Clustering Algorithm***: In clustering, the grouping of a particular set of objects based on their characteristics, according to their similarities takes place. Formation of cluster, a group of the same or similar elements gathered closely together, takes place. In this methodology, a collection of objects into a partition or nested set of partitions takes place. Hundreds of clustering algorithms are available, so are selected according to the optimal requirement.
- ***Classification Algorithm***: Classification algorithm is a procedure of selecting a hypothesis from a set of alternatives that best fits a set of observation. For data analysis, it is a technique used to predict group membership for data instances. This is used to analyze a given data set and takes each instance of it. It assigns this instance to a particular class. Such that classification error will be least. It is used to extract models. That defines

 ISSN: 2454-4248
101 – 104

important data classes within the given data set. It is a two step process. In the first step, the model created by applying classification algorithm. In the second step, the extracted model is tested against a predefined test data set. It is to measure the model trained performance and accuracy.

- ***Frequent Pattern Mining Algorithm***: This algorithm aimed at unsheathing frequent patterns in data in order to deduce knowledge that may help in decision making. Frequency pattern mining has applications ranging from intrusion detection and market basket analysis. Scalable parallel algorithms hold the key to addressing pattern mining in the context of big data.

VI. Tools For Big Data Processing

A large number of tools are there for Big Data Processing. Batch processing, interactive analysis and stream processing are the major technologies in which most of the processing tools work. Some current techniques for analysis of Big Data are being discussed here.

- ***Apache Hadoop and MapReduce****:* Apache Hadoop and MapReduce is the most established software platform for big data analysis. Map reduce is a programming model based on divide and conquer method for processing large datasets. The divide and conquer method is implemented in two steps such as Map step and Reduce Step. Hadoop works on two kinds of nodes such as master node and worker node. The master node divides the input into smaller sub problems and then distributes them to worker nodes in map step. Thereafter the master node combines the outputs for all the sub problems in reduce step.
- ***Apache Mahout****:* It aims at providing commercial and scalable machine learning techniques for large scale and intelligent data analysis applications. Clustering, classification, regression, pattern mining, evolutionary algorithms, and batch based collaborative filtering are the core algorithms of mahout.
- ***Apache Spark****:* It is an open source framework built for speed processing, and sophisticated analytics of big data processing.
- ***Dryad****:* Another popular programming model is Dryad, which is for implementation of parallel and distributed programs. This is basically for handling large context bases on dataflow graph. This approach consists of a cluster of computing nodes, and resources of a computer cluster are used by the user to run their program in a distributed way. Thousands of machines, each of them with multiple processors could be used by user. The main advantage of this process is that users are not supposed to know anything about concurrent programming.
- ***Storm****:* It is a fault tolerant and distributed real time computation system for processing large streaming data. It is easy to set up and operate, scalable, fault-tolerant to provide competitive performances. The processing of storm cluster is apparently similar to hadoop cluster. Here users run different topologies for different storm tasks. The computational technology is partitioned and distributed to a number of worker processes and then each worker process implements a part of the topology.
- ***Jaspersoft****: It is* open source software that produces reports from database columns. It is platform for analysis of big data and has fast data visualisation capacity on popular storage platforms. It can quickly explore big data without extraction, transformation, and loading.

VII. Real World Example of Big Data Analytics

- Using Big Data Analytics to boost Customer acquisition and retention. Using big data allows businesses to observe a number of customer related trends and patterns, which leads to trigger loyalty and maximize the data at your disposal.
- Using Big Data Analytics to solve advertiser's problem and offer marketing insights. Studying this, helps in matching consumers' expectations, changing product line and hence help in marketing and advertising technologies.
- Big Data Analytics for risk management. A risk management plan is a critical investment for any business. Big data analysts help to fore see a potential risk and mitigating it before it occurs.
- Big Data Analytics as a driver of innovations and product development. It helps the companies to innovate and redevelop their products. Organizations correct as much as data a possible before designing new products and re-designing the existing products.
- Use of Big Data in supply chain management. PepsiCo is a consumer packaged goods company that relies on huge volume of data for an efficient supply chain management. The clients of the company provide various reports including the warehouse inventory, data is then used for reconciling and forecasting the production and shipment needs.

VIII. Challenges and Issues

In order to move beyond the existing techniques and strategies used for machine learning and data analytics, some challenges need to be overcome. Some of them are as follows:

- Volume of the data which is increasing day by day with such a high pace. As the volume of data increases, the value of different data records decreases in proportion to type and quantity among other factors.
- In order to select an adequate method or design, a solid scientific foundation needs to be developed.
- Storage and transport issues also a very important role, as there has no storage medium for such huge data explosion especially through social media.

 ISSN: 2454-4248
101 – 104

- New efficient and scalable algorithms need to be developed.
- For proper implementation of devised solutions, appropriate development skills and technological platforms must be identified and developed.
- Security and privacy is most important challenge with big data as it consists of very sensitive, personal and crucial information related to the users which can't be shared publically.
- Issue related to data processing is also a major challenge for the analysis of the big data. As a minute mistake in the processing of data could result to disaster for the net outcome of the big data.
- Lack of Big Data professionals is a big challenge for the big data analytics.

IX. TSUGGESTIONS FOR FUTURE WORK

The data collected from different sources all over the world had widely increased or can say just doubled in amount in just 2 years of time interval. Without proper analysis of that data, the data can't be utilized as such. In order to resolve this, the development of techniques took place which could be facilitating the big data analysis. The development of powerful computers is a plus point for the implementation of these techniques which leads to automated systems. Data transformation to knowledge is not an easy task without utilizing high performance large-scale data processing and data mining processes. Many different systems such as fuzzy sets, rough sets, soft sets, neural networks, their generalizations and hybrid models obtained by combining two or more of these models have been found to be fruitful in representing data. More often, big data can be reduced to include only the important characteristics necessary from a particular study or objective depending upon the area of application. Often missing values are there in the data collected for study. These values are supposed to be generated or elimination of such tuples from the data set takes place before the process of analysis. The later approach sometimes leads to information loss and hence not preferred. This brings up many research issues in the industry and research community in forms of capturing and accessing data effectively. In addition to fast processing with high performance and high throughput, storing it efficiently for future use is another issue. Additionally, machine learning concepts and tools are gaining popularity among researchers to facilitate meaningful results through these concepts. Research in the area of machine learning for big data has focused on data processing, algorithm implementation, and optimization. Many of the machine learning tools for big data are started recently needs drastic change to adopt it.

X. CONCLUSION

Big Data is not just about collection of huge data, it is actually a concept of providing an opportunity to find new insight into existing data and analysis your future data. In recent years the speed of data generation is increasing drastically. For common people deducing results on the basis of analysis of such data is not an easy task. As a result, here various research issues, challenges, tools, methodologies used for big data analysts have been discussed. From this, it clear that every big data platform has its individual focus. Some of them are designed keeping in mind the batch processing system whereas some are based upon real-time analytic. Each big data platform has their specific functionalities which can be used as per situation. Different techniques used for the analysis include data mining, machine learning, statistical analysis, intelligent analysis, cloud computing, data stream processing, etc. It is believed that future researchers will pay more attention to these techniques to solve problems of big data effectively and efficiently to get better and accurate results.

REFERENCES

[1] Gandomi and M. Haider (2015). Beyond the hype: Big data concepts, methods, and analytics, *International Journal of Information Management*, 35(2), pp.137-144.

[2] K. Kambatla, G. Kollias, V. Kumar and A. Gram (2014). Trends in big data analytics, *Journal of Parallel and Distributed Computing*, 74(7), pp. 2561-2573.

[3] T. K. Das and P. M. Kumar (2013). Big data analytics: A framework for unstructured data analysis, *International Journal of Engineering and Technology*, 5(1), pp.153-156.

[4] D. P. Acharjya and Kauser Ahmed P (2016). A Survey on Big Data Analytics: Challenges, Open Research Issues and Tools, *International Journal of Advanced Computer Science and Applications*, 7(2), pp.511-519.

[5] Komal (2018). A Review Paper on Big Data Analytics Tools, *International Journal of Technical Innovation in Modern Engineering & Science*, 4(5), pp. 1012-1017.

[6] M. Chen, S. Mao, and Y. Liu (2014). Big data: a survey, *Mobile Networks and Applications*, 19(2), pp. 171–209.

[7] T. Erl, W. Khattak, and P. Buhler (2015). Big Data Fundamentals: Concepts, Drivers & Techniques, *Prentice Hall, India*, pp. 65-88.

[8] Hugh J. Watson (2014). Tutorial: Big Data Analytics: Concepts, Technologies and Applications, *Communications of the Association for Information Systems*, 34(65), pp. 1247-1268.

[9] Nana Kwame Gyamfi, Prince Appiah, Kofi Adu-Manu Sarpong, Silas Kwabla Gah, Ferdinand Katsriku and Jamal-Deen Abdulai (2017). Big Data Analytics: Survey Paper, *Conference Proceedings: Dialogue on Sustainability and Environmental Management, Accra, Ghana*, 15-16.

[10] Samiddha Mukherjee and Ravi Shaw (2016). Big Data – Concepts, Applications, Challenges and Future Scope *International Journal of Advanced Research in Computer and Communication Engineering*, 5(2), pp. 66-74.